Geostatistical Data Fusion for Remote Sensing Applications

Amy Braverman, Hai Nguyen, Edward Olsen, and Charles Miller
Jet Propulsion Laboratory, California Institute of Technology
Mail Stop 306-463
4800 Oak Grove Drive
Pasadena, CA 91109-8099
email: Amy.Braverman@jpl.nasa.gov

Noel Cressie, Matthias Katzfuss, and Rui Wang
Department of Statistics, The Ohio State University
1958 Neil Ave, 404 Cockins Hall
Columbus, OH 43210

Anna Michalak
Department of Civil and Environmental Engineering, University of Michigan
183 EWRE Building
Ann Arbor, MI 48109

***Abstract*—Remote sensing data are often sparse relative to the space-time domains of geophysical processes: no instrument observes everywhere, all the time. Remote sensing data are also massive, taken over different, usually non-nested spatial footprints, and subject to measurement error biases. Our goal is to infer the true values of a spatially and temporally continuous geophysical quantity from these aggregated, noisy, and heterogeneous observations. We do this through a geostatistical model that relates the observations to the true but not directly observed variable of interest. This model accounts for spatial correlations, and relationships among different resolutions. Crucially, our estimates are accompanied by uncertainty measures so we know how reliable the estimates are. In this talk, we review the basic principles of our methodology, as they apply to estimating column carbon dioxide in the planetary boundary layer, inferred from remote sensing observations taken at multiple scales with differing sensitivities and sampling characteristics.**

***Index Terms*—Spatial statistical inference, change of support, massive datasets, satellite remote sensing, carbon dioxide.**

I. Introduction

THIS article describes our work to date developing methodology for geostatistical (spatial statistical) data fusion. We regard data fusion as an inference problem: given two heterogenous input datasets with different statistical characteristics, how do we optimally estimate the quantity of interest, and obtain uncertainty measures associated with these inferences? Spatial statistics [1] provides a rigorous formalism for modeling spatial data: both the unknown true quantities of interest and the observations are described by random variables, and a statistical model specifies relationships among these variables. Spatial statistical models quantify relationships between observations and the true "processes" they measure in a way that exploits their spatial dependence. Importantly, these models also make it possible to describe relationships among quantities that exist at different spatial resolutions. In statistics this is called the change of support problem because the "support" of a random variable is the spatial unit to which it applies. Spatial statistical model provide a mechanism for making the desired inferences.

Spatial statistics is a special case of space-time statistics. The basic concepts were originally developed in two areas: geostatistics and time series analysis. Geostatistics developed around problems in mining applications and provides a set of techniques for making inferences that exploit spatial dependence. Time series analysis exploits temporal dependence in making inferences from observations over time. These problems share a common mathematical basis. The only differences between them are 1) dimensionality: time series are one dimensional, spatial problems are one, two, or three dimensional, and space-time problems are four dimensional; and 2) directionality: the fact that time only flows forward while it is possible to move both forward and backward in space.

The work described here builds on past, present, and ongoing research in space-time statistics. A key goal of our efforts is to find computational and mathematical compromises required to bring the power of this elegant formalism to bear on NASA's immense remote sensing data holdings. A recent methodological breakthrough called Fixed-Rank Kriging (FRK; [2]) now makes this possible. Traditional geostatistical methods have not been able to handle massive amounts of data because they require inverting very large matrices, or make tenuous assumptions about spatial dependence in order to avoid that requirement.

We demonstrate our methodology by applying it to the problem of estimating carbon dioxide (CO2) in the planetary boundary layer (PBL) of Earth's atmosphere using data from the Atmospheric Infrared Sounder (AIRS), and synthetic data representative of what the Orbiting Carbon Observatory (OCO-2) will see. PBL CO2 is an important quantity scientifically because the planetary boundary layer is that part of the atmosphere most immediately and directly affected by the flux of CO2 between the surface and the atmosphere. Fluxes should be correlated with changes in PBL CO2 amounts. Satellite observations from OCO and AIRS offer an unprecedented opportunity to quantify PBL CO2 globally and systematically, but only

This research is performed at the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

if their data can be properly combined.

This article is organized as follows. In the next section we provide some background on spatial statistics. This is followed by a somewhat more technical discussion of kriging and Fixed-Rank Kriging, and then some details about how we can use our model to correct for measurement biases and different supports in the observations. In Section VI we describe the extension of FRK to full spatial statistical data fusion, and in Section VII present the results of our fusion of AIRS and synthetic OCO. Finally, we conclude with a critique and discussion of next steps.

II. Background

Remote sensing data are by their very nature, statistical. Satellite instruments do not directly observe geophysical variables, they observe radiances from which geophysical information is inferred. Observations from space are typically made on spatial units coincident with instruments' pixels, while the true physical process is continuous in space. This discretization of the scene is one source of uncertainty. Another source is that the instrument itself adds measurement error at the pixel level. This error includes both bias ("systematic" error) and variance ("random" error). Finally, there may be additional bias and variability due to the inability of the instrument to observe under certain conditions (e.g. clouds). These relationships are illustrated in Figure 1.

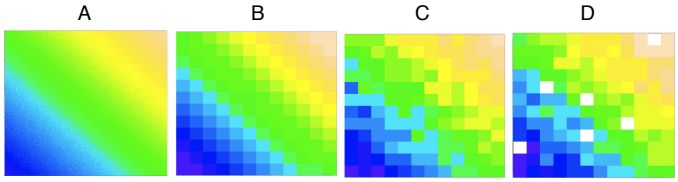

Fig. 1. Remote sensing data and their relationship to the true quantity of interest. The left panel shows the true geophysical field (A). (B) shows the field as it is viewed by a remote sensing instrument. The image is pixelated because each pixel is the average of the true values in (A) belonging to it. The instrument has measurement error, so (C) shows the image corrupted by noise. Noise is assumed to be independent from pixel to pixel. Finally, some pixels may not be observed at all due to instrument observing characteristics (e.g., some instruments can't see through clouds). (D) shows the image under these conditions.

The problem is that we only have access to information like that in panel D in Figure 1, and we want to infer the true continuous field in panel A. We can do this using a spatial statistical model that relates the observations in panel D to the true field in panel A.

Inference from spatial data requires a spatial statistical model. The spatial model is no different than any other statistical model except that the variables in it pertain to specific locations rather than people or other objects. Consider a simple linear regression model relating people's heights to their weights. If height and weight are assumed to be jointly normally distributed, then a regression analysis exploits the estimated correlation between these two variables. The optimal estimate of the weight of a randomly selected individual from the population used to derive the model is the mean of the conditional distribution of weight for the new person's height. The uncertainty of that estimate is the standard deviation of the conditional distribution.

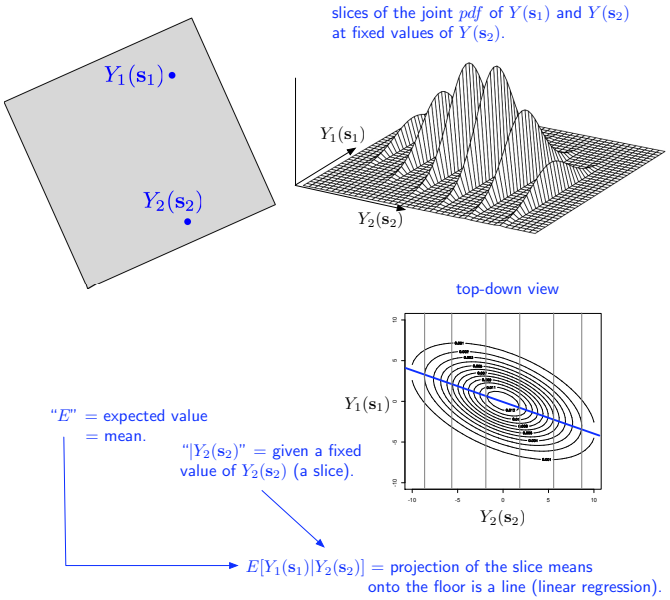

Fig. 2. Making inferences by exploiting spatial correlation.

The same principle applies in a spatial context, as illustrated in Figure 2. The top left graphic shows a square spatial domain with two locations, $\mathbf{s}_1$ and $\mathbf{s}_2$, identified by the blue dots. The true values of the geophysical quantity of interest are $Y_1(\mathbf{s}_1)$ and $Y_2(\mathbf{s}_2)$ where Y_1 and Y_2 may represent either two different underlying processes or the same underlying process ($Y_1 = Y_2 = Y$). The top right graphic shows the conditional distributions of $Y_1(\mathbf{s}_1)$ for specific values of $Y_2(\mathbf{s}_2)$ under the assumption that $Y_1(\mathbf{s}_1)$ and $Y_2(\mathbf{s}_2)$ are jointly bivariate gaussian. The top-down view of that joint distribution is shown in the graphic on the bottom right. The blue line is the regression line, which runs through the means of the conditional distributions in the center panel. If we know what happened at $\mathbf{s}_2$, and its correlation with $\mathbf{s}_1$, we can estimate the value of $Y_1(\mathbf{s}_1)$ and attach an uncertainty to that estimate. In fact, if we knew the values of Y_2 at multiple locations in the domain, say M locations, that would allow us to make better inferences. The joint distribution in Figure 2 would then be $(M+1)$-dimensional, and the conditional distribution used to estimate Y_1 at the location of interest would be an M-dimensional slice.

Remote sensing data pose two complications for the implementation of this estimation procedure. First, generally we do not observe at point-level. Satellite instruments observe averages of Y over pixel-size areas called footprints. Second, satellite instruments have measurement error which can impart bias and additional variability. These

relationships are expressed in the equation

$$Z(B(\mathbf{s})) = \left[\frac{1}{|B(\mathbf{s})|}\int_{\mathbf{u}\in B(\mathbf{s})} Y(\mathbf{u})d\mathbf{u}\right] + \epsilon(B(\mathbf{s})), \qquad (1)$$

where $B(\mathbf{s})$ is the instrument's footprint centered at location $\mathbf{s}$, $Z(B(\mathbf{s}))$ is the observed value over that footprint, $|B(\mathbf{s})|$ is the size of the footprint, and $\epsilon(B(\mathbf{s}))$ is the measurement error. Note that the measurement error term attaches to the footprint. Equation (1) relates the observed value, Z, to the true values from which it is constructed, $Y(\mathbf{s})$, $\mathbf{s} \in B(\mathbf{s})$, and ϵ. Inferences about Y's can therefore be made on the basis of Z if we have or assume some information about the characteristics of ϵ. The situation is similar to that depicted in the multivariate version in Figure 2, except that the axes corresponding to the predictors (the $Y_2(\mathbf{s}_2)$ axis in Figure 2) are a set of Z values corresponding to observations. This is shown in Figure 3. The predictand is still $Y(\mathbf{s}_1)$ since our objective is to make an inference about Y at a point location. The fact that the observations are aggregates, and the additional variation due to the variance of ϵ cause the variances of the conditional distributions to increase. Nonetheless, there is still a linear statistical relationship between $Y_1(\mathbf{s})$ and the set of Z's owing to the correlations among these quantities, that can be exploited.

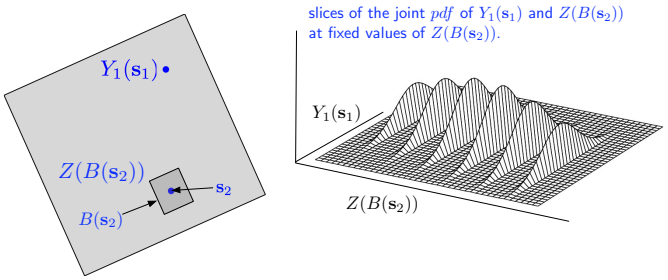

Fig. 3. Making inferences by exploiting spatial correlation between a point location ($\mathbf{s}_1$) and a block-level measurement ($Z(B(\mathbf{s}_2))$). $B(\mathbf{s})_2)$ is the block-level footprint centered at point location $\mathbf{s}_2$, and $Z(B(\mathbf{s}_2))$ is the data value associated with that footprint.

The optimal linear estimator (unbiased and minimum variance) of $Y_1(\mathbf{s}_1)$ is derived in Chapter 3 of [1], and is called the ordinary kriging estimator. It is widely used in geostatistics, but for remote sensing applications, there are some aspects that are problematic. The derivation of the kriging estimator requires the inversion of the covariance matrix describing the relationship between the observed data points, and of the covariance between the quantity of interest, $Y(\mathbf{s})$, and the observations. Because remote sensing datasets are often massive, inversion of the covariance matrix of the Z's is difficult if not impossible. To get around this, usual practice is to assume that the covariance is isotropic (the same in all directions) which allows the covariance matrix to be simplified in a way that allows for easy inversion. However, isotropy is a very tenuous assumption for most geophysical processes. A second simplifying assumption is that the covariance is stationary, which legitimizes deriving and employing a single estimate of the covariance matrix for use everywhere. Assuming that the covariance structure is the same everywhere on the Earth is also tenuous. To overcome these problems, [2] propose specifying the spatial covariance matrix in a way that does not require these assumptions, and permits inversion of very large matrices.

III. Kriging and Fixed-rank Kriging

Kriging, first developed by Georges Matheron and D.G. Krige [1], belongs to the family of least squares optimal linear interpolators. We start with assumptions for the spatial data generation process. Let $\{Y(\mathbf{s}) : \mathbf{s} \in D \subset \Re^d\}$ be a hidden, real-valued spatial process, and let $\mathbf{Z}$ be a vector of N point-level observations. Each element of $\mathbf{Z}$ is generated as the sum of $Y(\cdot)$ and an independent error term, $\epsilon(\cdot)$:

$$Z(\mathbf{s}) = Y(\mathbf{s}) + \epsilon(\mathbf{s}) \ , \quad \mathbf{s} \in D. \qquad (2)$$

The error term $\epsilon(\cdot)$ is assumed to be a white noise process with mean zero, and finite $\mathrm{Var}(\epsilon(\mathbf{s})) = \sigma^2 v(\mathbf{s})$, where $\sigma > 0$. $v(\cdot)$ is assumed known, and allows for the possibility of non-constant variance over the domain D. The hidden process $Y(\cdot)$ is assumed to have a linear mean structure:

$$Y(\mathbf{s}) = \mathbf{t}(\mathbf{s})'\boldsymbol{\alpha} + \nu(\mathbf{s}) \ , \quad \mathbf{s} \in D. \qquad (3)$$

The first term on the right side of (3) accounts for the linear trend, where $\mathbf{t}(\cdot) = (t_1(\cdot), \ldots, t_p(\cdot))$ is vector of p known covariates, such as geographical coordinates or other independent variables. $\boldsymbol{\alpha}$, the vector of linear coefficients, is unknown, and will be estimated from the data. The process $\nu(\mathbf{s})$ is assumed to have mean zero and finite, non-zero variance with spatial covariance function

$$\mathrm{Cov}(\nu(\mathbf{u}), \nu(\mathbf{v})) = C(\mathbf{u}, \mathbf{v}); \quad \mathbf{u}, \mathbf{v} \in D. \qquad (4)$$

Combining the Equations (2) - (4), we have the following general linear mixed model:

$$\mathbf{Z} = \mathbf{T}\boldsymbol{\alpha} + \boldsymbol{\delta}, \ \ \boldsymbol{\delta} = \boldsymbol{\nu} + \boldsymbol{\epsilon}, \qquad (5)$$

where $\boldsymbol{\delta}, \boldsymbol{\nu}$, and $\boldsymbol{\epsilon}$ are vectors of length N representing the corresponding processes evaluated at the available locations. $\mathbf{T}$ is an $N \times p$ matrix of the covariates $(\mathbf{t}(\mathbf{s}_1)', \ldots, \mathbf{t}(\mathbf{s}_N)')'$. The error term, $\boldsymbol{\delta}$, is a combination of the measurement error, $\boldsymbol{\epsilon}$, and the spatial covariance term, $\boldsymbol{\nu}$. Assuming independence between $\boldsymbol{\epsilon}$ and $\boldsymbol{\nu}$, the covariance matrix $\boldsymbol{\Sigma} = \mathrm{Var}(\boldsymbol{\delta})$ is

$$\boldsymbol{\Sigma} = \mathbf{C} + \sigma^2\mathbf{V}, \qquad (6)$$

where $\mathbf{C} \equiv [C(\mathbf{s}_i, \mathbf{s}_j)]$ and $\mathbf{V} \equiv \mathrm{diag}(v(\mathbf{s}_1), \ldots, v(\mathbf{s}_N))$.

We are interested in the hidden true process $Y(\cdot)$ at a location $\mathbf{s}_0, \mathbf{s}_0 \in D$. Traditional kriging solves for the best linear unbiased predictor, $\hat{Y}(\mathbf{s}_0) = \mathbf{a}'\mathbf{Z}$, by minimizing

$$||Y(\mathbf{s}_0) - \hat{Y}(\mathbf{s}_0)||^2 = ||Y(\mathbf{s}_0) - \mathbf{a}'\mathbf{Z}||^2.$$

[1] gives the following formulas for estimating of $Y(\mathbf{s}_0)$ and its mean-squared prediction error:

$$\hat{Y}(\mathbf{s}_0) = \mathbf{t}(\mathbf{s}_0)\hat{\boldsymbol{\alpha}} + \mathbf{k}(\mathbf{s}_0)'(\mathbf{Z} - \mathbf{T}\hat{\boldsymbol{\alpha}}), \qquad (7)$$

$$\text{MSPE}(\hat{Y}(\mathbf{s}_0)) = \{C(\mathbf{s}_0,\mathbf{s}_0) - \mathbf{k}(\mathbf{s}_0)'\boldsymbol{\Sigma}\mathbf{k}(\mathbf{s}_0) + (\mathbf{t}(\mathbf{s}_0) - \mathbf{T}'\mathbf{k}(\mathbf{s}_0))'(\mathbf{T}'\boldsymbol{\Sigma}^{-1}\mathbf{T})^{-1}(\mathbf{t}(\mathbf{s}_0) - \mathbf{T}'\mathbf{k}(\mathbf{s}_0))\}^{-\frac{1}{2}}, \quad (8)$$

where

$$\boldsymbol{\alpha} = (\mathbf{T}'\boldsymbol{\Sigma}^{-1}\mathbf{T})^{-1}\mathbf{T}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}, \quad (9)$$

$$\mathbf{k}(\mathbf{s}_0) = \mathbf{c}(\mathbf{s}_0)'\boldsymbol{\Sigma}^{-1}, \quad (10)$$

and $\mathbf{c}(\mathbf{s}_0) \equiv (C(\mathbf{s}_0,\mathbf{s}_1),\ldots,C(\mathbf{s}_0,\mathbf{s}_N))'$ is the vector of covariances between the process at prediction location $\mathbf{s}_0$ and the process at the observed locations corresponding to the elements of data vector $\mathbf{Z}$.

Optimal kriging coefficients require $\boldsymbol{\Sigma}$ and $\mathbf{c}(\mathbf{s}_0)$, both of which in turn require knowledge of the covariance function, $C(\mathbf{s}_i,\mathbf{s}_j)$. This quantity must typically be estimated from the data. For small datasets, simplifying assumptions are introduced in order to estimate of $C(\cdot,\cdot)$ [1]. Two common assumptions are stationarity and isotropy. Under stationarity, the covariance between process values at two locations $\mathbf{s}_i$ and $\mathbf{s}_j$ is a function of the difference vector between them. That is,

$$\text{Cov}(\mathbf{Z}(\mathbf{s}_i),\mathbf{Z}(\mathbf{s}_j)) = \text{C}(\mathbf{s}_1 - \mathbf{s}_2).$$

If the process is isotropic, or uniform in all directions, we can express this as a function of the distance between two locations,

$$\text{Cov}(\mathbf{Z}(\mathbf{s}_i),\mathbf{Z}(\mathbf{s}_j)) = \text{C}(d), \qquad d = |\mathbf{s}_i - \mathbf{s}_j|.$$

Such simplifying assumptions are not suitable for remote sensing data [3]. Physical processes on a global scale tend to vary differently depending on geographical regions. In addition, traditional kriging does not scale well with data size because it requires inversion of the $N \times N$ covariance matrix $\boldsymbol{\Sigma}$, a procedure requiring $O(N^3)$ computations. Fixed-ranked kriging (FRK; [2]) provides both computational scalability, and flexibility in fitting a wide range of covariance functions without requiring restrictive assumptions.

Equations (7) - (10) show that computation of the kriging coefficients requires the inversion of the covariance matrix $\boldsymbol{\Sigma} = \mathbf{C} + \sigma^2\mathbf{V}$. Cressie and Johannesson [2] model $C(\mathbf{u},\mathbf{v})$ as

$$C(\mathbf{u},\mathbf{v}) = \mathbf{S}(\mathbf{u})'\mathbf{K}\mathbf{S}(\mathbf{v}); \ \mathbf{u},\mathbf{v} \in D, \quad (11)$$

for some positive-definite matrix $\mathbf{K}$. $\mathbf{S}(\mathbf{u})$ is the basis expansion of $\mathbf{u}$ into a fixed set of r (not necessarily orthogonal) scalar-valued basis functions, $S_j(\cdot)$. That is,

$$\mathbf{S}(\mathbf{u}) = (S_1(\mathbf{u}),\ldots,S_r(\mathbf{u}))'; \ \mathbf{u} \in D.$$

The notation $\mathbf{S}$ without an argument denotes the $r \times N$ matrix constructed by applying $\mathbf{S}(\cdot)$ to all the locations in the dataset: $\mathbf{S} \equiv (\mathbf{S}(\mathbf{s}_1),\ldots,\mathbf{S}(\mathbf{s}_N))$.

The covariance model that combines Equations (6) and (11) is nonnegative-definite, with $\nu(\mathbf{s}) = \mathbf{S}(\mathbf{s})'\boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is an r-dimensional vector with $\text{Var}(\boldsymbol{\eta}) = \mathbf{K}$. Consequently, the data model in (2) can be written as

$$Z(\mathbf{s}) = \mathbf{t}(\mathbf{s})'\boldsymbol{\alpha} + \mathbf{S}(\mathbf{s})'\boldsymbol{\eta} + \boldsymbol{\epsilon}(\mathbf{s}). \quad (12)$$

The model for $\nu(\cdot)$ is called the *spatial random effects* (SRE) model, and the full model with the linear mean structure in Equation (12) is called the *spatial mixed effects* (SME) model [2].

To invert $\boldsymbol{\Sigma} = \sigma^2\mathbf{V} + \mathbf{S}'\mathbf{K}\mathbf{S}$, observe that it can be thought of as a rank-r update to the $N \times N$ diagonal matrix, $\sigma^2\mathbf{V}$. Using the Sherman-Morrison-Woodbury formula [4],

$$\boldsymbol{\Sigma}^{-1} = (\sigma^2\mathbf{V})^{-1} - (\sigma^2\mathbf{V})^{-1}\mathbf{S}'\left(\mathbf{K}^{-1} + \mathbf{S}(\sigma^2\mathbf{V})^{-1}\mathbf{S}'\right)^{-1}\mathbf{S}'(\sigma^2\mathbf{V})^{-1}.$$

The inversion is exact for any covariance function in the spatial random effects class. The procedure requires inversion of $\mathbf{K}$ and $(\mathbf{K}^{-1} + \mathbf{S}'(\sigma^2\mathbf{V})^{-1}\mathbf{S})$, both of which are $r \times r$ matrices. The number of computations required to invert $\boldsymbol{\Sigma}$ is $O(Nr^2)$, and grows only linearly with N. Cressie and Johannesson combine the FRK covariance model with (7) and (8) to produce the FRK kriging predictors and estimates of their uncertainties:

$$\hat{Y}(\mathbf{s}_0) = \mathbf{t}(\mathbf{s}_0)\hat{\boldsymbol{\alpha}} + \mathbf{S}(\mathbf{s}_0)'\mathbf{K}\mathbf{S}(\mathbf{Z} - \mathbf{T}\hat{\boldsymbol{\alpha}}),$$

$$\text{MPSE}(\hat{Y}(\mathbf{s}_0)) = \{\mathbf{S}(\mathbf{s}_0)'\mathbf{K}\mathbf{S} - \mathbf{S}(\mathbf{s}_0)'\mathbf{K}\mathbf{S}\boldsymbol{\Sigma}^{-1}\mathbf{S}'\mathbf{K}\mathbf{S}(\mathbf{s}_0) + (\mathbf{t}(\mathbf{s}_0) - \mathbf{T}'\boldsymbol{\Sigma}^{-1}\mathbf{S}'\mathbf{K}\mathbf{S}(\mathbf{s}_0))'(\mathbf{T}'\boldsymbol{\Sigma}^{-1}\mathbf{T})^{-1}(\mathbf{t}(\mathbf{s}_0) - \mathbf{T}'\boldsymbol{\Sigma}^{-1}\mathbf{S}'\mathbf{K}\mathbf{S}(\mathbf{s}_0))\}^{-\frac{1}{2}}.$$

$\boldsymbol{\Sigma}$ and $\mathbf{K}$ must be estimated from the data.

IV. Estimating $\boldsymbol{\Sigma}$ and $\mathbf{K}$: Change of Support

The discussion in the previous section assumed that the true quantity of interest and the measurements are taken a specific point locations. That is, we observe at point locations $\mathbf{s}_i$, and wish to make an inference at point location $\mathbf{s}_0$. However, satellite data often correspond to areal regions called footprints, and the following modified data model would then be more appropriate:

$$\mathbf{Z} = (Z(B_1), Z(B_2), \ldots, Z(B_n))'$$

$$Z(B_i) = \frac{1}{|B_i|}\int_{\mathbf{u}\in B_i} Y(\mathbf{u})d\mathbf{u} + \epsilon(B_i), \quad (13)$$

where $B_i = B(\mathbf{s}_i)$ is the footprint to which the observation applies. This is the average of the underlying process over the spatial domain of B_i, plus a random error measurement term, $\epsilon(B_i)$ as shown in Equation (13). We seek an estimate of the point-level covariance function, $C(\mathbf{u},\mathbf{v}) = \mathbf{S}'(\mathbf{u})\mathbf{K}\mathbf{S}(\mathbf{v})$, which would allow us to predict to any level of aggregation. The crucial term here is $\mathbf{K}$, which we can estimate from an examination of the covariances among areal-level data.

Without loss of generality, assume that the process $Y(\cdot)$ has mean zero. The covariance between two observations with support B_k and B_l is

$$\text{Cov}(Z(B_k), Z(B_l)) = \text{Cov}\left(\frac{1}{|B_k|}\int_{\mathbf{u}\in B_k} Y(\mathbf{u})d\mathbf{u}, \ \frac{1}{|B_l|}\int_{\mathbf{v}\in B_l} Y(\mathbf{v})d\mathbf{v}\right), \quad (14)$$

where $|B_i|$ represents the area of footprint B_i. Since the measurement errors are assumed to be independent of each other and of the Y's, the covariance between footprints reduces to the expression in Equation (14). Using the FRK covariance structure,

$$\begin{aligned}\mathrm{Cov}(Z(B_k),&Z(B_l)) = \\ &\frac{1}{|B_k|}\frac{1}{|B_l|}\int_{\mathbf{u}\in B_k}\int_{\mathbf{v}\in B_l}\mathrm{Cov}\left(Y(\mathbf{u}),Y(\mathbf{v})\right)\, d\mathbf{u}\, d\mathbf{v}, \\ &= \frac{1}{|B_k|}\int_{\mathbf{u}\in B_k}\mathbf{S}(\mathbf{u})' d\mathbf{u}\ \mathbf{K}\quad \frac{1}{|B_l|}\int_{\mathbf{v}\in B_l}\mathbf{S}(\mathbf{v})\, d\mathbf{v}, \\ &= \widetilde{\mathbf{S}}(B_k)'\ \mathbf{K}\ \widetilde{\mathbf{S}}(B_l), \end{aligned} \tag{15}$$

where

$$\widetilde{\mathbf{S}}(B_i) = \left(\widetilde{S}_1(B_i), \widetilde{S}_2(B_i), \ldots, \widetilde{S}_r(B_i)\right),$$

$$\text{with} \qquad \widetilde{S}_j(B_i) = \frac{1}{|B_i|}\int_{\mathbf{u}\in B_i} S_j(\mathbf{u}) \quad d\mathbf{u}.$$

Equation (15) is important because it expresses the covariance between aggregated footprints, $\mathrm{Cov}(Z(B_i), Z(B_j))$, in terms of the point-level covariance function $C(\cdot,\cdot)$. Note that $\mathbf{K}$ was defined at the point level. That is, $\mathbf{S}(\mathbf{s}_1)\,\mathbf{K}\,\mathbf{S}(\mathbf{s}_2)'$ is defined as the covariance between two point locations $\mathbf{s}_1$ and $\mathbf{s}_2$. Equation (15) indicates that the empirical covariance matrix of the spatial aggregates $\mathbf{Z}$ can be written as the product $\widetilde{\mathbf{S}}(B_k)'\ \mathbf{K}\ \widetilde{\mathbf{S}}(B_l)$, where $\widetilde{\mathbf{S}}(B_k)$ is the set of basis functions $\{S_j(\cdot)\}$ averaged over footprint B_k. Integrating the basis functions over the pixel footprints allows us to compute $\widetilde{\mathbf{S}}$, and subsequently estimate $\mathbf{K}$.

To estimate $\mathbf{K}$ and σ, we use detrended data since non-detrended data, $\mathbf{Z}$, would force $\mathbf{K}$ to capture both trend and covariance thereby skewing the estimates. In the absence of prior knowledge about spatial covariance, we model the trend using the ordinary least squares estimator,

$$\hat{\boldsymbol{\alpha}} = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{Z}.$$

With an estimate of $\boldsymbol{\alpha}$, we can calculate detail residuals $\mathbf{D} = \mathbf{Z} - \mathbf{T}\hat{\boldsymbol{\alpha}}$, which reflect only the spatial random effects process and measurement errors. Note that the bold $\mathbf{D}$ denotes the vector of detail residuals, and we let $D(\mathbf{s}_j)$ denote the detail residual for location $\mathbf{s}_j$. (D, without an argument, is the domain of the dataset). Following [2], we partition the domain, D, into M subregions, called bins, with centers $\mathbf{u}_j,\ j \in 1,\ldots,M$, and $M << N$. The bin centers $\mathbf{u}_j$ should provide good coverage of D. Around bin center $\mathbf{u}_j$, we define a neighborhood $\mathbf{N}(\mathbf{u}_j)$ and a set of 0-1 weights,

$$w_{jk} = \begin{cases} 1 & \text{if } \mathbf{s}_k \in \mathbf{N}(\mathbf{u}_j) \\ 0 & \text{otherwise,} \end{cases}$$

where $k \in 1,\ldots,N$. Let $\mathbf{w}$ be an $N \times M$ matrix defined by $\mathbf{w} = (\mathbf{w}'_1,\ldots,\mathbf{w}'_M)$, where $\mathbf{w}_j$ is a row vector of length N, $\mathbf{w}_j = (w_{j1},\ldots,w_{jN})$. $\mathbf{w}_j$ is the binary vector for bin center $\mathbf{u}_j$, indicating which of the N observations fall within the neighborhood around $\mathbf{u}_j$.

[2] defines an $N \times N$ positive-definite empirical covariance matrix $\hat{\boldsymbol{\Sigma}} \equiv [\hat{\boldsymbol{\Sigma}}(\mathbf{u}_i, \mathbf{u}_j)]$, where

$$\hat{\boldsymbol{\Sigma}}(\mathbf{u}_i,\mathbf{u}_j) = \begin{cases} V_D(\mathbf{u}_j) & i = j \\ C_D(\mathbf{u}_i,\mathbf{u}_j) & i \neq j, \end{cases}$$

and

$$\begin{aligned} C_D(\mathbf{u}_i,\mathbf{u}_j) &= \\ &\sum_{k_1=1}^{N}\sum_{k_2=1}^{N} w_{ik_1}w_{jk_2}D(\mathbf{s}_{k_1})D(\mathbf{s}_{k_2})/(\mathbf{w}'_i\mathbf{1}_N)(\mathbf{w}'_j\mathbf{1}_N), \\ V_D(\mathbf{u}_j) &= \sum_{k=1}^{N} w_{jk}D(\mathbf{s}_k)^2/(\mathbf{w}'_i\mathbf{1}_N). \end{aligned}$$

$\mathbf{S}$ and $\mathbf{V}$ can be binned in a similar fasion. We define the binned versions as $\bar{\mathbf{S}} = \mathbf{S}\mathbf{w}$ and $\bar{\mathbf{V}} = \mathbf{V}\mathbf{w}$, respectively. Cressie and Johannesson estimate $\mathbf{K}$ by minimizing

$$\begin{aligned} ||\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}(\mathbf{K},\sigma^2)||_F &= \\ &\mathrm{tr}((\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}(\mathbf{K},\sigma^2))'(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}(\mathbf{K},\sigma^2))), \end{aligned} \tag{16}$$

where $||\cdot||_F$ is the Frobenius norm. The $\mathbf{K}$ that minimizes Equation (16), given known σ^2, is

$$\hat{\mathbf{K}} = \mathbf{R}^{-1}\mathbf{Q}'(\hat{\boldsymbol{\Sigma}} - \sigma^2\bar{\mathbf{V}})\mathbf{Q}(\mathbf{R}^{-1})',$$

where $\bar{\mathbf{S}} = \mathbf{Q}\mathbf{R}$, $\mathbf{Q}$ and $\mathbf{R}$ via the QR decomposition of the binned $\bar{\mathbf{S}}$. Measurement error variance can be estimated by minimizing the following quantity with respect to σ^2:

$$\sum_{j,k}\left((\hat{\boldsymbol{\Sigma}} - \mathbf{Q}\mathbf{Q}'\hat{\boldsymbol{\Sigma}}\mathbf{Q}\mathbf{Q}')_{jk} - \sigma^2(\bar{\mathbf{V}} - \mathbf{Q}\mathbf{Q}'\bar{\mathbf{V}}\mathbf{Q}\mathbf{Q}')_{jk}\right)^2,$$

where $\bar{\mathbf{V}}$ is the binned version of $\mathbf{V}$, and $\mathbf{Q}$ is the orthogonal matrix derived from QR decomposition of $\bar{\mathbf{S}}$. $(\hat{\boldsymbol{\Sigma}} - \mathbf{Q}\mathbf{Q}'\hat{\boldsymbol{\Sigma}}\mathbf{Q}\mathbf{Q}')_{jk}$ indicates the element in the jth row and kth column of the matrix $(\hat{\boldsymbol{\Sigma}} - \mathbf{Q}\mathbf{Q}'\hat{\boldsymbol{\Sigma}}\mathbf{Q}\mathbf{Q}')$ [2]. Note that this is simply linear regression with slope σ^2, and intercept zero.

Finally, the covariance between the spatial aggregates in $\mathbf{Z}$ and the process at point location $\mathbf{s}$, $\mathrm{Cov}(\mathbf{Z}, Y(\mathbf{s})) = \mathbf{c}_\mathbf{s}$, is an N-vector with elements,

$$\begin{aligned}\mathrm{Cov}(Z(B_i),Y(\mathbf{s})) &= \frac{1}{|B_i|}\int_{\mathbf{u}\in B_i}\mathbf{S}_j(\mathbf{u})\ \mathbf{K}\ \mathbf{S}(\mathbf{s})'\ \ d\mathbf{u}\ , \\ &= \widetilde{\mathbf{S}}(B_k)'\ \mathbf{K}\ \mathbf{S}(\mathbf{s}).\end{aligned}$$

V. Correcting for Bias

As with any imperfect measurement, the potential exists in satellite observations for bias. If the bias is known, we can correct for it in the solution to the kriging equation. Assume that bias is multiplicative:

$$\begin{aligned}\mathrm{E}(\epsilon(\mathbf{s})) &= c\mu, \\ \mathrm{Var}(\epsilon(B_j)) &= \sigma^2,\end{aligned}$$

where $\mu = \mathrm{E}(Y(\mathbf{s}))$. The bias parameter, c, is generally estimated by regressing satellite observations on coincident

validation data. We then solve the kriging equation by minimizing mean-squared prediction error with the additional unbiasedness contraint on the kriging coefficients, $\mathbf{a}$. Thus, we minimize

$$||Y(\mathbf{s}_0) - \hat{Y}(\mathbf{s}_0)||^2 = ||Y(\mathbf{s}_0) - \mathbf{a}'\mathbf{Z}||^2$$

subject to

$$\begin{aligned} E(Y(\mathbf{s})) = \mu = E(\mathbf{a}'\mathbf{Z}) = E(\hat{Y}(\mathbf{s})), \\ \mu = \mathbf{a}'\mathbf{1}_N(1+c)\mu \\ 0 = \mathbf{a}'\mathbf{1}_N(1+c) - 1. \end{aligned}$$

This system can be solved using the method of Lagrange multipliers. The kriging coefficients, $\mathbf{a}$, and the Lagrange multiplier, m, can be found by solving the matrix equation,

$$\begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{1}_N(1+c) \\ \mathbf{1}'_N(1+c) & 0 \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ m \end{bmatrix} = \begin{bmatrix} \mathbf{c}(s) \\ 1 \end{bmatrix}.$$

The solution will have the same form as (7) and (8), except that the kriging coefficients $\mathbf{a}$ will be different as a result of the constraint:

$$\begin{aligned} \mathbf{a} = &\left(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}\mathbf{1}_N(1+c)\left(-\mathbf{1}'_N(1+c)\boldsymbol{\Sigma}^{-1}\mathbf{1}_N(1+c)\right)^{-1}\right. \\ &\left.\mathbf{1}'_N(1+c)\boldsymbol{\Sigma}^{-1}\right)\mathbf{c}(\mathbf{s}_0) + \boldsymbol{\Sigma}^{-1}\mathbf{1}_N(1+c)\left(-\mathbf{1}'_N(1+c)\boldsymbol{\Sigma}^{-1}\right. \\ &\left.\mathbf{1}_N(1+c)\right)^{-1}\mathbf{1}'_N(1+c)\boldsymbol{\Sigma}^{-1}. \end{aligned}$$

VI. Spatial Statistical Data Fusion

The previous section discussed how to make optimal, point-level inferences from aggregated, footprint-level data. In this section, we describe a natural extension of that procedure for making optimal, point-level inferences from multiple footprint-level input data sets that may have heterogeneous geometries and other error characteristics. We call this Spatial Statistical Data Fusion (SSDF; or, interchangeably, geostatistical data fusion). First, however, we must clarify what we are trying to estimate.

In Section II we noted that the measurements taken at different locations may be measurements of the same quantity ($Y = Y_1 = Y_2$) or of different quantities ($Y_1 \neq Y_2$). Mathematically, the more general case is that $Y_1 \neq Y_2$, and we may be interested in simply inferring the pair $(Y_1(\mathbf{s}), Y_2(\mathbf{s}))$ at location $\mathbf{s}$, or inferring the value of some function of them, $g(Y_1(\mathbf{s}), Y_2(\mathbf{s}))$.

Let $\mathbf{Z}_1$ be a realization of process $Y_1(\cdot)$, and similarly let $\mathbf{Z}_2$ be a realization of process $Y_2(\cdot)$, with

$$\begin{aligned} \mathbf{Z}_i &= (Z_i(B_{i1}), Z_i(B_{i2}), \ldots, Z_i(B_{iN_i}))' \\ Z_i(B_{ij}) &= \frac{1}{|B_{ij}|}\int_{\mathbf{u}\in B_{ij}} Y(\mathbf{u})d\mathbf{u} + \epsilon_i(B_{ij}), \end{aligned}$$

where B_{ij} represents the jth footprint from dataset i. We also assume a multiplicative bias model:

$$\begin{aligned} \mathrm{E}(\epsilon_i(B_{ij})) &= c_i\mu, \\ \mathrm{Var}(\epsilon_i(B_{ij})) &= \sigma_i^2. \end{aligned}$$

At every point location $\mathbf{s}$, we estimate the vector of underlying processes,

$$\hat{\mathbf{Y}}(\mathbf{s}) = \begin{pmatrix} \hat{Y}_1(\mathbf{s}) \\ \hat{Y}_2(\mathbf{s}) \end{pmatrix} = \begin{pmatrix} \mathbf{a}'_{11\mathbf{s}}\mathbf{Z}_1 + \mathbf{a}'_{12\mathbf{s}}\mathbf{Z}_2 \\ \mathbf{a}'_{21\mathbf{s}}\mathbf{Z}_1 + \mathbf{a}'_{22\mathbf{s}}\mathbf{Z}_2 \end{pmatrix}. \quad (17)$$

The coefficient vectors $\mathbf{a}_{ij\mathbf{s}}$ are unknown. The subscript i in $\mathbf{a}_{ij\mathbf{s}}$ denotes the index of the process, the subscript j denotes the index of the dataset, and $\mathbf{s}$ denotes the location. We solve for the optimal coefficients by minimizing the mean-squared prediction error (MSPE) of $\hat{\mathbf{Y}}(\cdot)$:

$$\mathrm{E}\left[\left(\hat{Y}_1(\mathbf{s}) - Y_1(\mathbf{s})\right)^2 + \left(\hat{Y}_2(\mathbf{s}) - Y_2(\mathbf{s})\right)^2\right], \quad (18)$$

with the unbiasedness constraints,

$$\begin{aligned} \mathrm{E}(\hat{Y}_1(\mathbf{s})) &= \mathbf{a}'_{11\mathbf{s}}\mathbf{1}_{N_1}\mu_1 + \mathbf{a}'_{12\mathbf{s}}\mathbf{1}_{N_2}\mu_2 = \mu_1, \\ \mathrm{E}(\hat{Y}_2(\mathbf{s})) &= \mathbf{a}'_{21\mathbf{s}}\mathbf{1}_{N_1}\mu_1 + \mathbf{a}'_{22\mathbf{s}}\mathbf{1}_{N_2}\mu_2 = \mu_2, \end{aligned} \quad (19)$$

assuming, for the moment, that μ_1 and μ_2, are known. We will address this assumption in more detail below in Subsection B. The optimal fusion coefficients are obtained by forming the lagrangian from (18) and (19), and differentiating with respect to $\mathbf{a}_{11\mathbf{s}}$, $\mathbf{a}_{12\mathbf{s}}$, $\mathbf{a}_{21\mathbf{s}}$, and $\mathbf{a}_{22\mathbf{s}}$. The full derivation is given in [3], and the result is

$$\begin{aligned} \mathbf{a}_{i1s} &= \mathbf{A}_1^{-1}(\mathbf{B}_{i1} + \mathbf{C}_1 m), \\ \mathbf{a}_{i2s} &= \mathbf{A}_2^{-1}(\mathbf{B}_{i2} + \mathbf{C}_2 m). \end{aligned} \quad (20)$$

m is the lagrange multiplier:

$$m = \frac{(1 - \mathbf{1}'_{N_1}\mathbf{A}_1^{-1}\mathbf{B}_{i1}(1+c_1) + \mathbf{1}'_{N_2}\mathbf{A}_2^{-1}\mathbf{B}_{i2}(1+c_2))}{(\mathbf{1}'_{N_1}\mathbf{A}_1^{-1}\mathbf{C}_1(1+c_1) + \mathbf{1}'_{N_2}\mathbf{A}_2^{-1}\mathbf{C}_2(1+c_2))},$$

and

$$\begin{aligned} &\mathbf{A}_1 \equiv (\mathbf{I}_{N_1} - \check{\boldsymbol{\Sigma}}_{11}^{-1}\check{\boldsymbol{\Sigma}}_{12}\check{\boldsymbol{\Sigma}}_{22}^{-1}\check{\boldsymbol{\Sigma}}_{21}), \text{ an } N_1 \times N_1 \text{ matrix}, \\ &\mathbf{A}_2 \equiv (\mathbf{I}_{N_2} - \check{\boldsymbol{\Sigma}}_{22}^{-1}\check{\boldsymbol{\Sigma}}_{21}\check{\boldsymbol{\Sigma}}_{11}^{-1}\check{\boldsymbol{\Sigma}}_{12}), \text{ an } N_2 \times N_2 \text{ matrix}, \\ &\mathbf{B}_{i1} \equiv \boldsymbol{\Sigma}_{11}^{-1}(\mathbf{c}_{i1s} - \check{\boldsymbol{\Sigma}}_{12}\check{\boldsymbol{\Sigma}}_{22}^{-1}\mathbf{c}_{i2s}), \text{ an } N_1\text{-dimensional vector}, \\ &\mathbf{B}_{i2} \equiv \check{\boldsymbol{\Sigma}}_{22}^{-1}(\mathbf{c}_{i2s} - \check{\boldsymbol{\Sigma}}_{21}\check{\boldsymbol{\Sigma}}_{11}^{-1}\mathbf{c}_{i1s}), \text{ an } N_2\text{-dimensional vector}, \\ &\mathbf{C}_1 = -\check{\boldsymbol{\Sigma}}_{11}^{-1}(\mathbf{1}_{N_1}(1+c_1) - \check{\boldsymbol{\Sigma}}_{12}\check{\boldsymbol{\Sigma}}_{22}^{-1}\mathbf{1}_{N_2}(1+c_2)), \\ &\qquad\qquad \text{an } N_1\text{-dimensional vector}, \\ &\mathbf{C}_2 = -\check{\boldsymbol{\Sigma}}_{22}^{-1}(\mathbf{1}_{N_2}(1+c_2) - \check{\boldsymbol{\Sigma}}_{21}\check{\boldsymbol{\Sigma}}_{11}^{-1}\mathbf{1}_{N_1}(1+c_1)), \\ &\qquad\qquad \text{an } N_2\text{-dimensional vector}, \end{aligned}$$

where $\mathbf{c}_{ijs} = \mathrm{Cov}(Y_i(\mathbf{s}_0), \mathbf{Z}_j)$, $\mathbf{I}_{N_i}$ is the $N_i \times N_i$ identity matrix, and $\check{\boldsymbol{\Sigma}}_{ij}$ is given by the FRK covariance model,

$$\mathrm{Cov}(Z_i(B_{il}), Z_j(B_{jk})) = \mathbf{S}(B_{il})'\hat{\mathbf{K}}_{ij}\mathbf{S}(B_{jk}) + \hat{\sigma}^2\mathbf{V}. \quad (21)$$

Having solved for the optimal coefficient vectors $\mathbf{a}_{ij\mathbf{s}}$, we can estimate the vector of processes, $\mathbf{Y}(\mathbf{s})$, by substituting $\mathbf{a}_{ij\mathbf{s}}$ into Equation (17). The mean-squared prediction error for $\hat{Y}_i(\mathbf{s})$ may be obtained by

$$\begin{aligned} \mathrm{MPSE}(\hat{Y}_i(\mathbf{s})) = &\mathbf{a}'_{i1\mathbf{s}}\mathrm{Var}(\mathbf{Z}_1)\mathbf{a}_{i1\mathbf{s}} + \mathbf{a}'_{i2\mathbf{s}}\mathrm{Var}(\mathbf{Z}_2)\mathbf{a}_{i2\mathbf{s}} + \\ &\mathrm{Var}(Y_1(\mathbf{s})) + 2\mathbf{a}'_{1\mathbf{s}}\mathrm{Cov}(\mathbf{Z}_1, \mathbf{Z}_2)\mathbf{a}_{2\mathbf{s}} - \\ &2\mathbf{a}'_{1\mathbf{s}}\mathrm{Cov}(\mathbf{Z}_1, Y_1(\mathbf{s})) - 2\mathbf{a}'_{2\mathbf{s}}\mathrm{Cov}(\mathbf{Z}_2, Y_1(\mathbf{s})). \end{aligned} \quad (22)$$

There are several issues specific to the fusion of data from multiple processes that must be addressed. These include estimation of the matrix $\mathbf{K}$ and unbiasedness constraints. We discuss these considerations in the next two subsections.

A. Estimating $\boldsymbol{K}$ for two processes

Obtaining robust estimates of the empirical covariance matrices generally requires large amounts of data. In practice, having large datasets is crucial to our calculations, as the massive number of observations allows for robust and stable estimates of the cross-covariance matrices without relying on simplifying assumptions such as stationarity and isotropy. The derivations above do not specify a covariance model for $\mathbf{\Sigma}_{ij}$. We assume the FRK covariance model, since it possesses required scalability and change of support properties. However, with two processes to be estimated, we require three different $r \times r$ matrices to capture the covariance functions. Let $\mathbf{K}_{11}$ model the covariance in dataset 1, $\mathbf{K}_{22}$ the covariance in dataset 2, and $\mathbf{K}_{12}$ the cross-covariance between the two datasets. The relationships between the covariance matrices and their respective FRK models are:

$$\begin{aligned}\mathbf{\Sigma}_{11} &= \mathbf{S}_1'\mathbf{K}_{11}\mathbf{S}_1 + \sigma_1^2\mathbf{V}_1,\\ \mathbf{\Sigma}_{12} &= \mathbf{S}_1'\mathbf{K}_{12}\mathbf{S}_2,\\ \mathbf{\Sigma}_{21} &= \mathbf{S}_2'\mathbf{K}_{21}\mathbf{S}_1,\\ \mathbf{\Sigma}_{22} &= \mathbf{S}_2'\mathbf{K}_{22}\mathbf{S}_2 + \sigma_2^2\mathbf{V}_2.\end{aligned}$$

The parameters $\mathbf{K}_{ij}$ may be estimated by a procedure similar to the one described in Section IV.

In this section we assume that the data are observed at point support. It is not difficult to adjust two-process SSDF to accommodate areal-level input data. Using the FRK covariance structure, the covariance between the process at any two areal-level inputs, B_j and B_k, is

$$C(B_j, B_k) = \int_{\mathbf{u}\in B_j} \mathbf{S}(\mathbf{u})' d\mathbf{u} \mathbf{K} \int_{\mathbf{u}\in B_k} \mathbf{S}(\mathbf{u})' d\mathbf{u}.$$

Therefore, corrections for change of support can be incorporated by integrating the function $\mathbf{S}(\cdot)$ over the relevant footprints as in Section IV.

B. Unbiasedness constraints for two processes

Earlier we assumed that the mean parameters, μ_1 and μ_2, are known. In reality, this is usually not the case. Though it is possible to estimate them, the estimates are often unstable [1]. Alternatively, under certain assumptions, the parameters can be removed without estimation. See [3] for discussion. In general, however, without prior knowledge of μ_1 and μ_2, the unbiasedness constraint in (18) reduces to the following more general constraints:

$$\begin{aligned}\mathbf{a}_{i1\mathbf{s}}'\mathbf{1}_{N_1}\mu_1 &= \mu_1,\\ \mathbf{a}_{i1\mathbf{s}}'\mathbf{1}_{N_1} &= 1.\end{aligned} \tag{23}$$

and

$$\begin{aligned}\mathbf{a}_{i2\mathbf{s}}'\mathbf{1}_{N_2}\mu_2 &= \mu_2,\\ \mathbf{a}_{i2\mathbf{s}}'\mathbf{1}_{N_2} &= 1.\end{aligned} \tag{24}$$

Therefore, in the absence of knowledge about μ_1 and μ_2, we can maintain the unbiasedness constraints by adding four Lagrange multipliers terms instead of two. The derivation then parallels that leading to Equations (20) with one difference: inversion by block must be applied two extra times to solve for the additional Lagrange multipliers. (Again see [3] for details.)

The unbiasedness constraints given by (23) and (24) lead to a solution that amounts to ordinary co-kriging. Ordinary co-kriging is co-kriging with the assumption that the two datasets have constant and unknown means [1]. Note that this solution is not optimal for the special case where $Y_1(\cdot) = Y_2(\cdot)$, since it does not exploit the fact that $\mu_1 = \mu_2$.

Finally, if the bias characteristics of $\epsilon_i(B_{ij})$ are known, we can correct for measurement bias as we did in Section V. If the measurement error terms $\epsilon_i(B_{ij})$ have

$$\begin{aligned}\mathrm{E}(\epsilon_i(B_{ij})) &= c_i\mu_i,\\ \mathrm{Var}(\epsilon_i(B_{ij})) &= \sigma_i^2,\end{aligned}$$

then we correct for multiplicative biases by modifying (19) as follows:

$$\begin{aligned}\mathrm{E}(\hat{Y}_1(\mathbf{s})) &= c_1\mathbf{a}_{11\mathbf{s}}'\mathbf{1}_{N_1}\mu_1 + c_2\mathbf{a}_{12\mathbf{s}}'\mathbf{1}_{N_2}\mu_2 = \mu_1,\\ \mathrm{E}(\hat{Y}_2(\mathbf{s})) &= c_1\mathbf{a}_{21\mathbf{s}}'\mathbf{1}_{N_1}\mu_1 + c_2\mathbf{a}_{22\mathbf{s}}'\mathbf{1}_{N_2}\mu_2 = \mu_2.\end{aligned}$$

In the single dataset case in Section III the measurement error variance, σ_i^2, was estimated along with $\mathbf{K}$, and the bias, c_i was assumed known. Here, σ_i^2 could be estimated along with $\mathbf{K}_{ii}$ as before. However, if the biases are not known, they must be estimated by comparison of measurements to ground truth. In that case, an alternative is to estimate σ_i^2 from those validation data at the same time as we estimate bias.

C. Estimating a Linear Combination of Two Processes

In certain applications, it may be necessary to estimate a linear combination of multiple processes, where the weights are known, but the data for the processes are observed individually. Suppose we have two datasets, $\mathbf{Z}_1$ and $\mathbf{Z}_2$, generated from two process, $Y_1(\cdot)$ and $Y_2(\cdot)$. For location $\mathbf{s}$, we wish to construct an estimate of $Y_w(\mathbf{s}) = w_1Y_1(\mathbf{s}) + w_2Y_2(\mathbf{s})$, where the weights w_i are known. One approach might be to separately estimate $Y_1(\mathbf{s})$ and $Y_2(\mathbf{s})$, and combine them. This is, however, suboptimal, since it does not exploit the covariance between the two processes. Another approach is to construct a new dataset $\mathbf{W}(\mathbf{s}) = w_1Z_1(\mathbf{s}) + w_2Z_2(\mathbf{s})$ and estimate $Y_w(\cdot)$ from $\mathbf{W}$. This approach only works if the datasets are observed at the same locations and with identical footprints.

To estimate the process $Y_w(\mathbf{s})$ at location $\mathbf{s}$, we use

$$\hat{Y}_w(\mathbf{s}) = w_1\left(\mathbf{a}_{11\mathbf{s}}'\mathbf{Z}_1 + \mathbf{a}_{12\mathbf{s}}'\mathbf{Z}_2\right) + w_2\left(\mathbf{a}_{21\mathbf{s}}'\mathbf{Z}_1 + \mathbf{a}_{22\mathbf{s}}'\mathbf{Z}_2\right).$$

We use the machinery of Section VI to derive an 2×1 vector of estimates of the processes, $(\hat{Y}_1(\mathbf{s}), \hat{Y}_2(\mathbf{s}))'$, and an 2×2 mean-squared prediction error matrix, $\mathrm{Var}((\hat{Y}_1(\mathbf{s}), \hat{Y}_2(\mathbf{s}))' - (Y_1(\mathbf{s}), Y_2(\mathbf{s}))')$, at location $\mathbf{s}$. The estimate and the mean-squared prediction error for $Y_w(\mathbf{s})$

can be computed through simple matrix multiplication:

$$\hat{Y}_w(\mathbf{s}) = w_1\hat{Y}_1(\mathbf{s}) + w_2\hat{Y}_2(\mathbf{s}) = (w_1, w_2)\cdot\begin{pmatrix}\hat{Y}_1(\mathbf{s})\\ \hat{Y}_2(\mathbf{s})\end{pmatrix},$$

$$\mathrm{MPSE}(\hat{Y}_w(\mathbf{s})) = (w_1, w_2)\cdot\mathrm{Var}\left(\begin{pmatrix}\hat{Y}_1(\mathbf{s})\\ \hat{Y}_2(\mathbf{s})\end{pmatrix} - \begin{pmatrix}Y_1(\mathbf{s})\\ Y_2(\mathbf{s})\end{pmatrix}\right)\cdot(w_1, w_2)'.$$

VII. Fusing Carbon Dioxide Measurements from OCOL and AIRS

To illustrate the methods described above, consider estimating the difference between total column carbon dioxide (CO2) measured by an instrument like OCO (the Orbiting Carbon Observatory), and carbon dioxide in the mid-troposphere and above measured by the Atmospheric Infrared Sounder (AIRS). If both instruments had the same footprints and the same sampling characteristics, the difference between their measurements would be the amount of CO2 below the troposphere:

$$Y_w(\mathbf{s}) = Y_1(\mathbf{s}) - Y_2(\mathbf{s}).$$

This is an important quantity scientifically because it roughly corresponds to the amount of CO2 in the planetary boundary layer (PBL). The PBL is that portion of the atmosphere that is dragged along by the rotation of the Earth, and is most directly affected by flux of CO2 between the surface (land and ocean) and the atmosphere. Fluxes should be correlated with changes in PBL CO2 amounts. Satellite observations from OCO and AIRS offer an unprecedented opportunity to quantify PBL CO2 globally and systematically, but only if their data can be properly fused.

A. Input Data

Figure 4 is a schematic diagram of AIRS and OCO horizontal and vertical sampling. AIRS is on the Aqua spacecraft, and observes CO2 in the mid-troposphere and above on 90 km footprints. The AIRS sensitivity to different parts of the atmosphere is illustrated by the blue curve in the lower-right plot. AIRS data south of 60°S have not been vetted by the AIRS CO2 team, and so we exclude those values. To estimate AIRS measurement error characteristics, we repeated a portion of the work performed by the AIRS CO2 Validation Team, and compared AIRS retrievals coincident with in-situ aircraft flask measurements in the mid-troposphere. We found that AIRS is unbiased and has measurement error variance $(3.43)^2$ [5].

Due to the OCO launch failure, we use synthetic OCO data ("OCO-Like"; OCOL) in place of actual OCO data. We used geolocation information from the AIRS Level 2 Standard Product to compute the centers of footprints at which OCO would have observed. These locations correspond to an along-track strip of four, 1.1×2.25 km pixels down the middle of the center footprint of each AIRS granule. We assigned times 15 minutes ahead of the times associated with the AIRS center footprints, and gave all OCOL footprints cloud fractions equal to the cloud fraction of that center AIRS footprint. We then simulated OCOL measurements for pixels with cloud fraction less than .50, and only in daylight. The simulated values are downscaled from coarse resolution Parallel Climate Transport Model (PCTM) output as described in [6]. The OCO vertical sensitivity, shown by the red curve in the lower-right of Figure 4, was captured prior to downscaling by summing PCTM CO2 values over 28 vertical levels using the same pressure weighting scheme as would have applied for OCO. Finally, we assumed measurement error had no bias and variance of 2.25 ppm based on previous experience of our science advisors who have worked with PCTM for some time.

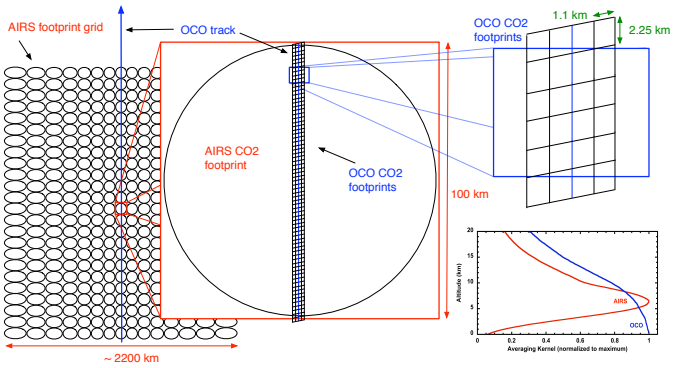

Fig. 4. Schematic diagram of AIRS and OCO horizontal and vertical sampling.

There are two important things to note about Figure 4. First, there is a vast spatial mismatch between footprints from the two instruments. A naive strategy would be to average all OCO footprints belonging to a given AIRS footprint and declare that the AIRS and average OCO values are commensurate, but this completely ignores the individual strengths of the two missions' designs, and does not account for the fact that the OCO footprints are only comparable to the center of the AIRS track.

AIRS has a coarse footprint, but near-global coverage every day. OCO has a very fine footprint, but poor daily coverage. For SSDF, these differences are an advantage rather than a disadvantage: we exploit the complementary nature of the two instruments' sampling to produce fused estimates that acknowledge spatial variability and achieve good coverage. Temporal mismatch is not particularly extreme here- just 15 minutes, but we are currently working to incorporate time into SSDF based on new work by Cressie and collaborators [7]. This will be important for applications where there are substantial temporal differences between observation times.

The second important thing to note in Figure 4 is that the vertical sensitivities of the two instruments (shown in the lower-right of the figure) are fairly similar in and above the mid-troposphere. In fact, neither instrument is uniformly sensitive there; both sensitivities decay rapidly with altitude. If the target of our inference is uniformly

weighted CO2 concentration, one could plausibly make the argument that both OCO and AIRS are biased measurements of this quantity, though they are biased in different ways owing to the difference between their sensitivities in the PBL. The nature of those biases is quantified by the expected values of the measurement error terms, ϵ_i, which would have to be estimated from ground truth information. We have defined our target not as total column CO2, but as PBL CO2 thus avoiding the need to quantify biases relative to a single, unobserved quantity (uniformly weighted total column CO2). Instead, we regard OCO and AIRS measurements as biased observations of two different quantities, a function of which (the difference) is the object of interest.

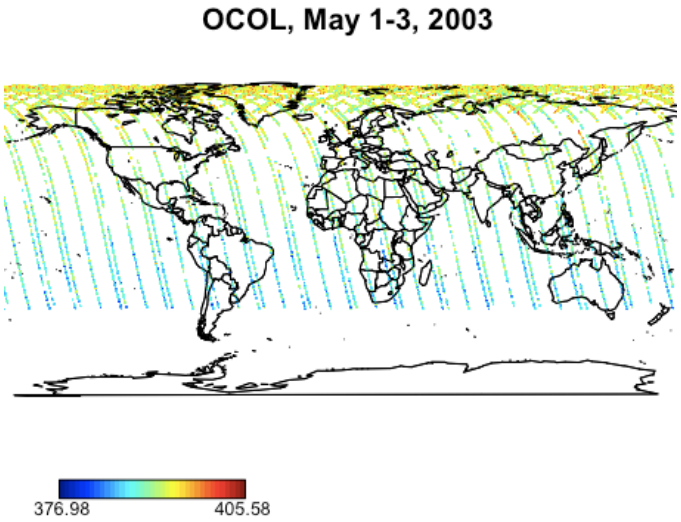

Fig. 5. Synthetic OCO observations of total column CO2 for May 1-3, 2003. The color scale represents parts per million (ppm).

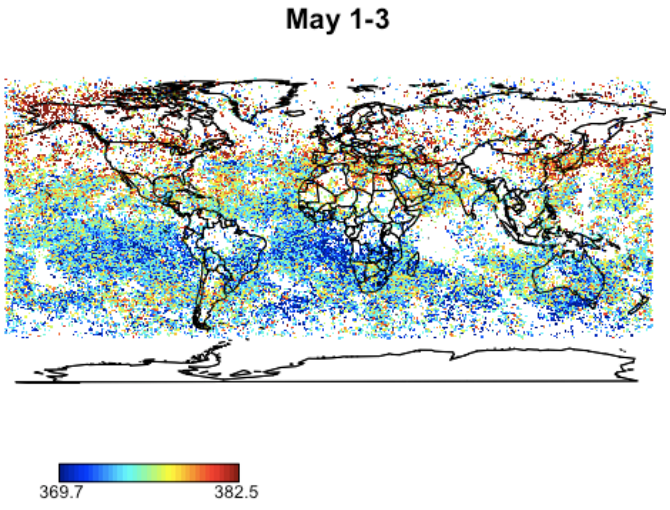

Fig. 6. Actual AIRS observations of CO2 in the mid-troposphere and above for May 1-3, 2003. The color scale represents parts per million (ppm).

We applied the methodology described in Section VI using $r = 396$ bisquare basis functions [3] to 89, overlapping, three day blocks of data from AIRS and OCOL for the period May 1, 2003 to July 31, 2003. The raw data for the first three day block, May 1-3, 2003, are shown in Figures 5 and 6. The second three day block is May 2-4, and so on. Besides the region south of 60°S which was deliberately excluded, AIRS has fairly dense coverage, but with several prominent gaps. OCOL presents a regular pattern with swaths that pass through but do not completely cover the AIRS gaps.

B. Fusion Results

Figures 7 and 8 show the results of fusing AIRS and OCOL for the three days May 1-3, 2003. Figure 7 shows the optimal estimates of PBL CO2 at each location of a $1° \times 1°$ grid. That is, we chose a $180° \times 360°$ grid of locations, $\mathbf{s}$ at which to apply our algorithm. Figure 8 shows the corresponding $\sqrt{\text{MPSE}}$ for each estimate. The color scales of the maps have been truncated at the maximum values of 25 ppm and 0.5 ppm respectively to avoid skewed color scales that would be caused by a few outliers.

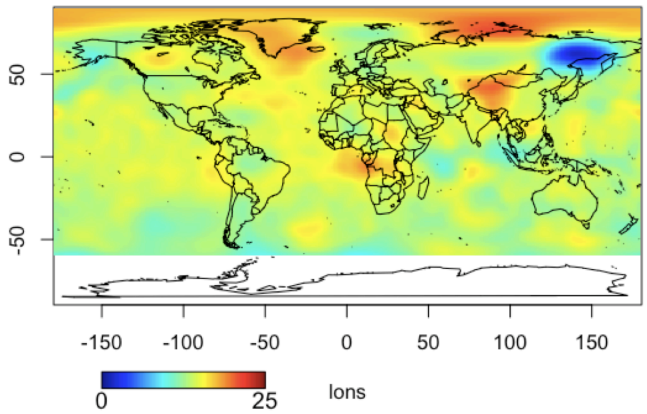

Fig. 7. Fused estimates of PBL CO2 for May 1-3, 2003. Estimates are made at the centers of one degree grid cells. The color scale represents parts per million (ppm).

Several features in the two maps are worth mentioning. First, in northern hemisphere summer we expect low CO2 values in the northern hemisphere due to the CO2 draw-down when vegetation foliates. Second, there is a pronounced area of low PBL CO2 over eastern Siberia, but these estimates also have relatively high uncertainties (values of $\sqrt{\text{MPSE}}$). There is also an area of relatively high PBL CO2, also with high uncertainty, over western China. In general areas where AIRS was sparse have higher uncertainties than areas where AIRS was dense. This is to be expected, since in these areas the estimation procedure relies more heavily on a combination of spatial interpolation of AIRS and the small amounts of OCOL present. This effect will be less where there is strong spatial correlation in the AIRS data. Finally, we could have estimated PBL CO2 in the region south of 60°S. Since there is no data there, estimates would have been based exclusively

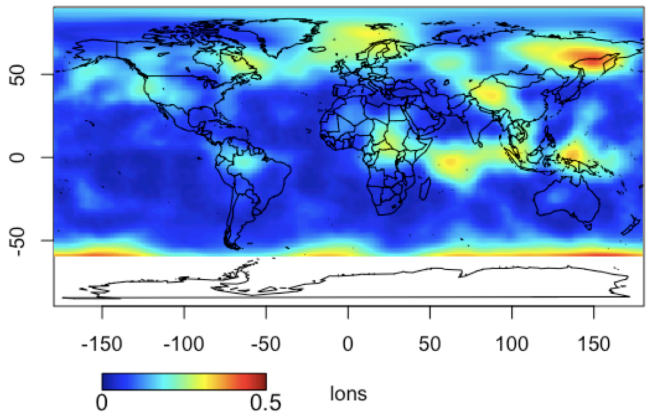

Fig. 8. Square root of the mean-squared prediction error for the estimates in Figure 7. The color scale represents parts per million (ppm).

on data in the bordering areas, and on the strength of spatial correlations between those data and the target region. Uncertainties would be high for such extrapolations- we can see that already in the higher values of the square root mean-squared prediction errors just below 50°S.

C. Computation

The analysis discussed in this article was carried out on a MacBook Pro laptop computer with a dual core, 3 GHz Intel processor. The computation time for each three day block of data was approximately five minutes, of which roughly half was consumed by the computation of mean-squared prediction errors for the estimates. We used $r = 396$ bisquare basis functions at three different levels of resolution, and a $1^\circ \times 1^\circ$ grid of estimation locations. This was chosen because 1° is approximately the size of an AIRS footprint (at the equator). Predictions on finer grids are certainly possible, but would be more computationally intensive.

VIII. Summary and Conclusions

We have produced fusion estimates of planetary boundary layer CO2 and their associated uncertainties on a one degree, spatial grid from three months of actual AIRS CO2 retrievals and synthetic OCO based on climate model output. We chose to break these data into 89, overlapping three day blocks to smooth out transitions from one block to the next. Three days is enough time to obtain a fairly complete dataset, and is also a short enough period that it can be regarded as a snapshot of current conditions. We made various choices such as the resolution of the output grid, the number of basis functions, the methods of estimating key parameters, etc. which can be debated, but our goal was to demonstrate the viability of our approach.

The obvious outstanding question is how do we validate these results? How do we know if our estimates are accurate? In principle, our uncertainties provide a measure of the average (squared) differences between our estimates and the true values, but many assumptions went into the calculations of both the estimates themselves and those uncertainties. We can perform various tests to check those assumptions (see [3]), but the gold standard would be to obtain independent, ground truth measurements of PBL CO2 at a representative set of locations and times. Such data are available only for a very limited set of cases for PBL CO2 (one location at Park Falls, WI at present), and this will not change significantly in the near future. Instrument validation teams routinely rely on such sparse ground truth sources because they have no choice but to use what is available.

An alternative validation strategy is to work entirely with synthetic data. That is, to downscale model output to fine resolution, then aggregate this information to mimic the resolutions of the instruments, and add synthetic measurement error. Then, we could fuse the synthetic observations to obtain estimates and their uncertainties and compare these results to the downscaled model data. We had to do that already for OCO, and will now move to this strategy in general by doing the same thing for AIRS. This will allow us to study the behavior of our method under different circumstances, and quantify performance tradeoffs for different algorithmic choices.

Our next development efforts will be devoted to incorporating time into this methodology. Existing work on Fixed-Rank Filtering [7] forms the basis of spatio-temporal data fusion (STDF). We expect this to provide an improvement over the naive strategy of simply breaking the data into fixed blocks. The main question will be the computational cost and feasibility of such an approach.

In closing, we note that this data fusion methodology and algorithm have wide applications to remote sensing data. We have concentrated thus far on the fusion of geophysical products, but there is no reason why one would not want to fuse Level 1 radiance products prior to performing joint retrievals, for example. This would be preferable to naive "match-up" as a way of merging data sets with different spatial, temporal, and other statistical characteristics. Another important application area is the combination of two instrument records that may or may not overlap in time. This will be the case for AIRS, OCO, and ASCENDS when ASCENDS starts to provide information later this decade. A number of NASA's Decadal Survey missions could potentially benefit from our fusion methodology, either because they inherently involve multiple kinds of observations, or because they are intended to extend existing data records.

References

[1] N. A. Cressie, *Statistics for Spatial Data*, John Wiley and Sons, Inc., 1993.

[2] N. Cressie and G. Johannesson, "Fixed rank kriging for spatio-temporal data," *Journal of the Royal Statistical Society: Series B*, vol. 70, no. 1, pp. 209–226, 2002.

[3] H. Nguyen, *Spatial Statistical Data Fusion for Remote Sensing Applications*, Ph.D. thesis, University of California, Los Angeles, 2009.

[4] M. A. Woodbury, "The stability of out-input matrices," *Mathematical Reviews*, 1949.

[5] A. Braverman N. Cressie M. Katzfuss H. Nguyen E. Olsen, "Characterization of AIRS Total Measurement Error for Data Fusion," unpublished manuscript available on eBooks, 2009.
[6] M. Katzfuss and N. Cressie, "Simulation of OCO-like Data based on PCTM Data," unpublished manuscript available on eBooks, 2010.
[7] E. Kang N. Cressie and T. Shi, "Fixed rank filtering for very large spatial data sets," *Journal of Computational and Graphical Statistics*, 2010.